the reality of ‘digital science’

kaitlin thaney
SciPy, 13 july 2011
austin, texas
xi. background
about me
Digital Science
(the company)
investment arm
incubator role
in-house dev
tiered approach
build to scale
researcher-focused
1. science, tech, and moving online
blocking points
(to name a few ... )

idea

publish

lit review
discovery

materials

share results

analyze

retest

experiment

collect data

(to name a few ... )
discovery & delivery
changes at the workbench
annotation & curation
social &
administrative
gaps still exist

"I think you should be more explicit here in step two."
2. key constituencies
machines
researchers
decision makers
machines

researchers
decision makers
“I do not fear computers. I fear the lack of them.”

— Isaac Asimov

annotation
markup
search
discovery
“behind the scenes”
[brief interlude]
digitisation of the scholarly canon

(content is still king)
PHILOSOPHICAL
TRANSACTIONS:
GIVING SOME
ACCOUNT
OF THE PRESENT
Undertakings, Studies, and Labours
OF THE
INGENIOUS
IN MANY
CONSIDERABLE PARTS
OF THE
WORLD.

Vol I.
For Anno 1665, and 1666.

In the SAVOY,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane,
Printers to the Royal Society.
not nearly there yet ...
Identification of Positionally Distinct Astrocyte Subtypes whose Identities Are Specified by a Homeodomain Code

Christian Hochstüm, Benjamin Deneen, Agnes Lukaszewicz, Qiao Zhou, and David J. Anderson

Abstract

Astrocytes constitute the most abundant cell type in the central nervous system (CNS) and play diverse functional roles, but the ontogenetic origins of this phenotypic diversity are poorly understood. We have investigated whether positional identity, a fundamental organizing principle governing the generation of neuronal subtype diversity, is also relevant to astrocyte diversification. We identified three positionally distinct subtypes of white-matter astrocytes (WMA) in the spinal cord, which can be distinguished by the combinatorial expression of Reelin and Slit1. These astrocyte subtypes derive from progenitor domains expressing the homeodomain transcription factors Pax6 and Nkx6.1, respectively. Loss- and gain-of-function experiments indicate that the positional identity of these astrocyte subtypes is controlled by Pax6 and Nkx6.1 in a combinatorial manner. Thus, positional identity is an organizing principle underlying astrocyte, as well as neuronal, subtype diversification and is controlled by a homeodomain.
barriers to “access”
still the starting point
patents are no better
(in many cases, worse)
SureChem is making patent chemistry search faster, easier and more accessible. SureChem indexes the full text of patent documents, delivers new structures within 24 hours of patents being issued, and gives users unprecedented access to valuable chemical structure data.

Our SureChem Portal enables users to search US, EP and WO patents along with MEDLINE and Japan patent abstracts quickly and cost-effectively and export both structure and patent results to their desktop.

SureChem’s unique Web Service and Database products enable researchers to perform batch screens analyses of proprietary compounds against the patent chemistry landscape, all in-house.

Find out about features and pricing here or contact us for more information.
New Chemical Structures Extracted From Patents This Week

Latest Patents From USPTO Applications For Top Pharma Firms

<table>
<thead>
<tr>
<th>Company</th>
<th>Patent Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pfizer</td>
<td>20100249128</td>
<td>Novel 3-AMIDO-PYRROLO[3,4-C]PYRAZOLE-5(1H,4H,6H) CARBALDEHYDE DERIVATIVES</td>
</tr>
<tr>
<td>GlaxoSmithKline</td>
<td>20100249137</td>
<td>Novel sEH Inhibitors and Their Use</td>
</tr>
<tr>
<td>Novartis</td>
<td>20100249107</td>
<td>Biomarkers for Alzheimer's Disease Progression</td>
</tr>
<tr>
<td>Sanofi-Aventis</td>
<td>20100248224</td>
<td>DNA FRAGMENTATION ASSAY</td>
</tr>
<tr>
<td>AstraZeneca</td>
<td>20100247617</td>
<td>Transdermal Administration Of (2S)-(4E)-N-Methyl-5-(3-(5-Iso propy lpyridinyl)-4-Pentan-2-Amine</td>
</tr>
<tr>
<td>Hoffmann-La Roche</td>
<td>20100230284</td>
<td>REAGENT DEVICES</td>
</tr>
<tr>
<td>Merck &amp; Co.</td>
<td>20100243956</td>
<td>LIQUID-CRYSTALLINE MEDIUM AND LIQUID-CRYSTAL DISPLAY</td>
</tr>
<tr>
<td>Abbott Laboratories</td>
<td>20100247576</td>
<td>NUCLEOTIDE AND AMINO ACID SEQUENCES ENCODING AN EXPORTED PROTEIN 1 DERIVED FROM PLASMODIUM VIVAX AND USES THEREOF</td>
</tr>
<tr>
<td>Eli Lilly and Co.</td>
<td>20100249075</td>
<td>SUBSTITUTED TETRALINS AS SELECTIVE ESTROGEN RECEPTOR-BETA AGONISTS</td>
</tr>
</tbody>
</table>

SureChem Database Statistics

Summary:

- Total Patents: 18,876,320
- Total Unique Structures: 11,758,454
- Last Update: Tuesday 30 November

Dataset Breakdown:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Documents</th>
<th>Unique Structures</th>
</tr>
</thead>
<tbody>
<tr>
<td>USPTO Granted</td>
<td>4,312,920</td>
<td>8,408,720</td>
</tr>
<tr>
<td>USPTO Applications</td>
<td>2,543,818</td>
<td>6,724,978</td>
</tr>
<tr>
<td>WO/PCT</td>
<td>1,425,433</td>
<td>5,478,016</td>
</tr>
<tr>
<td>European Applications</td>
<td>1,573,182</td>
<td>3,130,885</td>
</tr>
<tr>
<td>European Granted</td>
<td>895,125</td>
<td>2,872,392</td>
</tr>
<tr>
<td>MedLine</td>
<td>18,546,931</td>
<td>524,386</td>
</tr>
<tr>
<td>Japanese Abstracts</td>
<td>8,125,842</td>
<td>242,543</td>
</tr>
</tbody>
</table>
New Chemical Structures Extracted From Patents This Week

Latest Patents From USPTO Applications For Top Pharma Firms

- Pfizer
- GlaxoSmithKline
- Novartis
- Sanofi-Aventis
- AstraZeneca
- Hoffmann-La Roche
- Merck & Co.
- Abbott Laboratories
- Eli Lilly and Company

SureChem Databases

- Chemical Structure
- Basic Text
- Advanced Text
- Patent Number

Structure search and drawing powered by ChemAxon
New Chemical Structures Extracted From Patents This Week

Latest Patents From USPTO Applications For Top Pharma Firms

Pfizer
GlaxoSmithKline
Novartis
Sanofi-Aventis
AstraZeneca
Hoffmann-La Roche
Merck & Co.
Abbott Laboratories
Eli Lilly and Company

SureChem Databases

Structure Results 1 To 10 Of 1350 For

1. SMILES: OC(C)(O)OC(C)OC(C)(O)OC1N1=O=O=O=COC1=COC1=O
   Name: 10-farnesyl-4,8,8-dihydroxy-5,10-dihydrobenzo[b][1,4]diazepin-11-one
   Title: Benzodiazepine analogues. processes for their production and their use as pharmacuticals

EXAMPLE 21
3-(4-hydroxy-3-methoxyphenyl)-8-{[3-hydroxy-1-pyrdinyl](carbonyl)5,10-dihydro-11H-dibenzo[b,e]1,4]diazepin-11-one

EXAMPLE 23
3-(4-hydroxy-3-methoxyphenyl)-8-{[2-hydroxymethyl-1-piperidinyl](carbonyl)5,10-dihydro-11H-dibenzo[b,e]1,4]diazepin-11-one

The desired product was prepared by substituting 2-piperidinylmethanol for 3-pyrrolidin-1-ylpropylamine in Example 14. MS (DCI) m/e 474 (M+H)+; 'H NMR (500 MHz, DMSO-d6) 6 8.84 (s, 1H), 9.29 (brs, 1H), 8.08 (s, 1H), 7.72 (d, J=8.3 Hz, 1H), 7.25 (d, J=1.5 Hz, 1H), 7.16-7.21 (m, 2H), 7.09 (m, 1H), 6.97-7.03 (m, 2H), 6.88 (d, J=8.0 Hz, 1H), 4.35 (m, 1H), 3.86 (s, 3H), 3.28-3.34 (m, 4H), 1.48-1.80 (m, 6H).

EXAMPLE 24

The desired product was prepared by substituting ethyl 2-piperidinocarboxylate for 3-pyrrolidin-1-ylpropylamine in Example 14. MS (DCI) m/e 516 (M+H)+; 'H NMR (500 MHz, DMSO-d6) 6 9.89 (s, 1H), 9.28 (s, 1H), 8.14 (s, 1H), 7.73-7.74 (d, J=8.2 Hz, 1H), 7.24 (d, J=1.2 Hz, 1H), 7.17-7.19 (m, 2H), 7.07-7.09 (dd, J=8, 2.1 Hz, 1H), 6.95-7.05 (m, 3H), 6.88 (d, J=8.2 Hz, 1H), 4.16 (brs, 2H), 3.86 (s, 3H), 3.61 (m, 1H), 1.29-1.79 (m, 11H.).
can streamline
name disambiguation
10,11-dihydro-5-methyl-5H-dibenzo[b,e][1,4]diazepin-11-one

*(still strains the minds of the best)*
machine readability is key.

agreement is hard.
10,11-dihydro-5-methyl-5H-dibenzo[b,e][1,4]diazepin-11-one
“everything is metadata ...  
everything can be a label.”

- david weinberger
EVERYTHING IS DEEPLY INTERTWINGLED.
(Carbethoxymethylene)triphenylphosphorane (1.57 g; 4.5 mM) was added in a thin stream to a 50 ml reaction vessel containing 60% ethyl acetate/hexane (2107 cm\(^{-1}\) (N3); FAB-MS: MH\(^+\)=336; 1HNMR (CDCl3): 7.04 (s, 1H), 6.94 (d, 1H, J=7Hz), 4.09 (m, 1H), 2.4 (m,2H), 1.91 (s, 3H), 1.28 (t, 3H, J=7Hz).

**EXAMPLE 8: 5'-Carbethoxymethyl-3'-amino-5',3'-dideoxothymidine (18)**
A solution of the ester of Example 7 (16, 700 mg, 2.09 mM) in methanol (15 ml) was subjected to hydrogenation in a Parr apparatus (30 psi) in the presence of 10% Pd/C (210 mg; 30 wt%).

The hydrogenation reaction was completed after 24 hours. The catalyst was filtered and the filtrate evaporated to give 620 mg (96%) of the title compound. Rf=0.3 (5% saturated. NH3 in MeOH in EtOAc). FAB-MS: MH\(^+\)=312.

1HNMR (CDCl3): 7.4 (s,1H), 6.11 (dd, 1H, J=5Hz, 3Hz), 4.1 (q,2H, J=7Hz), 3.62 (m, 1H), 2.51-2.10 (m, 7H), 1.87 (s, 3H), 1.24 (t, 3H, J=7Hz) ppm.

**EXAMPLE 9: 5'-Carbethoxymethylene-3'-amino-5',3'-dideoxothymidine (10)**

Triphenylphosphine (786 mg; 3 mM) was added to a stirred solution of the ester of Example 7 (16, 700 mg; 2.09 mM) in THF/H2O (10 ml/l ml) and the reaction was monitored by gas evolution via an attached gas bubbler. The reaction was completed after 3 hours. The mixture was evaporated in vacuo and the crude oil was chromatographed over SiO\(_2\) (50 g). eluting with 5% saturated ammonium.

The term [[thiaoxyalkoxyalkyl]]; as used herein refers to R 80 S-R 81 O- wherein R 80 is loweralkyl as defined above and R 81 is alkylene.

Representative examples of alkoxyalkoxy groups include CH 3 SCH 2 O-, EtSCH 2 O-, t-BuSCH 2 O- and the like.

The term [[thiaoxyalkoxyalkyl]]; as used herein refers to a thiaoxyalkoxy group appended to an alkyl radical. Representative examples of alkoxyalkoxyalkyl groups include CH 3 SCH 2 CH 2 OCH 2 CH 2 -, CH 3 SCH 2 OCH 2 CH 2 -, and the like.

The term [[trans,trans]] as used herein refers to the orientation of substituents (R 1 and R 2) relative to the central substituent R as shown.

The term [[trans,cis]] as used herein refers to the orientation of substituents (R 1 and R 2) relative to the central substituent R as shown.

This definition encompasses both the case where R 1 are cis and R and R 1 are trans and the case where R 2 and R are trans and R and R 1 are cis.

The term [[cis,cis]] as used herein refers to the orientation of substituents (R 1 and R 2) relative to the central substituent R as shown.

Preferred compounds of the invention are selected from the group consisting of:

- trans,trans-2-(4-Methoxyphenyl)-4-(1,3-benzodioxol-5-yl)-1-[3-(N-propyl-N-n-pentanesulfonfylamino)propyl]-pyrrolidine-3-carboxylic acid;
- trans,trans-2-(4-Methoxyphenyl)-4-(1,3-benzodioxol-5-yl)-1-[(N-propyl-N-n-pentanesulfonfylamino)ethyl]pyrrolidine-3-carboxylic acid;
- trans,trans-2-(3,4-Dimethoxyphenyl)-4-(1,3-benzodioxol-5-yl)-1-[(N-propyl-N-n-pentanesulfonfylamino)ethyl]pyrrolidine-3-carboxylic acid;
- trans,trans-2-(3,4-Dimethoxyphenyl)-4-(1,3-benzodioxol-5-yl)-1-[(N-propyl-N-n-hexanesulfonfylamino)ethyl]pyrrolidine-3-carboxylic acid;
(Carbethoxymethylene)triphenylphosphorane (1.57 g; 4.5 mM) was added in a portion to a solution of the azide in water (20 ml) and extracted into ethyl acetate (2×50 ml), washed with brine and dried over anhydrous sodium sulfate. After removal of the solvent, the residue was purified by flash chromatography (hexanes: 60% ethyl acetate) to afford 765 mg of the title compound as a white solid (77% yield following chromatographic purification, two steps).

**Syntheses of the 17-keto precursors to cortistatins J, K and L.** The 17-keto cortistatin J precursor (32) was synthesized in just three steps from the azido alcohol 5. Following reductive (di)methylation to form the dimethylamino alcohol 30 (85% yield), 1,6-elimination of water and deprotection of the tert-butyldimethylsilyl ether occurred simultaneously upon stirring the dimethylamino alcohol 30 with a biphasic mixture of chloroform and concentrated hydrochloric acid (20 min, 23 °C). Without purification, the product of the latter transformation (31) was subjected to oxidation with the Dess–Martin periodinane, providing the 17-keto cortistatin J precursor (32) as a white solid (77% yield following chromatographic purification, two steps).
machines
researchers
decision makers
a few edge cases

(though no field is perfect)
A search for excited leptons in pp Collisions at $\sqrt{s} = 7$ TeV

CMS Collaboration

Subjects: High Energy Physics - Experiment (hep-ex)

A search for excited leptons is carried out with the CMS detector at the LHC, using 36.1 fb$^{-1}$ of pp collisions at $\sqrt{s} = 7$ TeV. A significant excess of events above the standard model expectation is observed. Interpreting the findings in the framework of new physics, lower limits on the mass of the excited lepton are reported. For $\Lambda = M^*$, excited leptons are discovered at a mass of 133 GeV.
tracking expiration calibration
ordering + processing
protocols
parameters
calibration
misc. lit
managing information
different types of “data”
often gigabytes, not terabytes
“i invented a folder based system ...”
“i invented a folder based system ...”

“yeah, we had a LIMS. it only ever got used to store photos from lab nights out.”
why?
experimentation reliance
data moves, grows legs
funder/instit’n pressure
Organize your Research.

BioKM is a web-based research management service that ensures your research is on the right track.

Welcome to a New Way of Research Management

Click To Start
30 day free trial

Researcher

Meet your research goals
Ensure you are on the right research track. Communicate your research progress. Stay focused to meet your research goals.

Principal Investigator

Cut down on research costs
Provide your team the right set of tools to perform better science, make sure research context is kept and the assets of your efforts are maintained.

Lab Manager

Ensure continuity of research
We know you are doing your best that everything will run smooth in the lab, combating lab entropy. BioKM was created to reduce the mundane tasks, providing you a tool for...
## Antibody Search Result

**Search result for:**

Huntingtin

<table>
<thead>
<tr>
<th>Sort by</th>
<th>Rating</th>
<th>Company</th>
<th>Keyword</th>
<th>Catalogue number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Huntingtin Antibody</td>
<td></td>
<td></td>
<td></td>
<td>$ - N/A</td>
</tr>
<tr>
<td>Host</td>
<td>Mouse</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Application</td>
<td>Polyclonal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>0.05ml</td>
<td></td>
<td></td>
<td>$ - N/A</td>
</tr>
<tr>
<td>Reacts With</td>
<td>HU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>QC</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publications</td>
<td>No</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Huntingtin Antibody (2E10)**

- Company: Novus Biologicals
- Catalogue #: H00003064-M01
- Clonality: Monoclonal
- Size: 0.1mg

- Antigen: Huntingtin
- Host: Mouse
- Reacts With: HU
- QC: No
- Publications: No
machines
researchers
decision makers
rewards, incentives the “why”
data capture
(of a different sort)
the **“social issue”**

best practices

behaviour roadblocks

discipline / researcher specific
WONKA'S GOLDEN TICKET

GREETINGS TO YOU, THE LUCKY FINDER OF THIS GOLDEN TICKET FROM MR. WILLY WONKA...

PRESENT THIS TICKET AT THE FACTORY GATES AT TEN O'CLOCK IN THE MORNING OF THE FIRST DAY OF OCTOBER AND DO NOT BE LATE. YOU MAY BRING WITH YOU ONE MEMBER OF YOUR OWN FAMILY... AND ONLY ONE... BUT NO ONE ELSE......

In your wildest dreams you could not imagine the marvelous surprises that await YOU.
imperfect system
“Right now we're going through a Cambrian explosion of metrics.”

- Johan Bollen
there’s been a drastic spike in terms of sheer **volume** and **type**
citation / impact factor

\( h \)-index

weighted citations (eigenfactor, sjr)

“betweenness centrality”

alt-metrics, etc.
difficult to ... harmonise track / maintain / map understand (even still measure)
administrators / funders = part of the research cycle
Dedicated to the needs of higher education

Symplectic's software solutions have been developed with leading academic institutions and specifically designed with their needs in mind.

They make day-to-day information management easier for research managers, administrators and researchers, and put the data you need for long-term strategic planning at your fingertips.

Symplectic's products can integrate seamlessly with your existing systems and all come with comprehensive technical support and training.

Symplectic Elements - publications management system

Our flagship product was used by Imperial College London to revolutionise the way it
Headline statistics

- Total number of people in selected groups: 1
- Total number of publications in selected groups: 18 (of which 18 are distinct)

Charts

**Key**

- HOOK, Daniel W (own)

**Average publications per year**

- Year 2006: 3
- Year 2008: 5
- Year 2010: 2

**Citations histogram**

- 1-5 citations: 8
- 6-10 citations: 6
- 11-15 citations: 4
- 16-20 citations: 2

**H-Index**

- H-index: 6

**Total publications by year**

<table>
<thead>
<tr>
<th>Year</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>2005</td>
<td>2</td>
</tr>
<tr>
<td>2006</td>
<td>3</td>
</tr>
<tr>
<td>2007</td>
<td>3</td>
</tr>
<tr>
<td>2008</td>
<td>4</td>
</tr>
<tr>
<td>2009</td>
<td>4</td>
</tr>
<tr>
<td>2010</td>
<td>2</td>
</tr>
</tbody>
</table>

**Journals (by frequency)**

<table>
<thead>
<tr>
<th>Journal</th>
<th>Impact factor</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>J PHYS AMATH THEOR</td>
<td>1.540</td>
<td>0</td>
</tr>
<tr>
<td>PHYS REV LETT</td>
<td>7.180</td>
<td>1</td>
</tr>
<tr>
<td>PRAMANA-J PHYS</td>
<td>0.274</td>
<td>1</td>
</tr>
</tbody>
</table>
3. the reality
“the future is here ... just not evenly distributed yet.”

- William Gibson
changing understandings, paradigms
technology can help design decisions are key plan for the irrational
more efficient research
increase productivity
enable reproducibility
thank you.

k.thaney@digital-science.com
www.digital-science.com
@kaythaney